Insensibility of Question Word Order in Visual Question Answering

Hwanhee Lee[1] and Kyomin Jung[1,2]
[1]Seoul National University, Seoul, Korea
[2]Automation and models Research Institute, Seoul National University, Seoul, Korea
{wanted1007,kjung}@snu.ac.kr

Abstract

In this paper we demonstrate whether the word order of input question affects the performance of existing Visual Question Answering(VQA) model trained on VQA v2 dataset. We show that randomly shuffling the word order of input question has little influence on the performance of VQA model. Also, we demonstrate that encoding the question with self-attention architecture that is independent of word order, shows nearly equal performance to Recurrent Neural Networks(RNNs) based encoder in existing VQA model. Based on these results, we show that existing VQA model does not utilize the word order of input question.

1 Introduction

Visual question answering(Goyal et al., 2017) is a task of answering a question related to a given image. The basic VQA model architecture fuses the question representation from RNNs and the image representations from Convolutional Neural Networks(CNNs) with attention mechanism.

Although there have been notable attention mechanisms (Ben-younes et al., 2017; Kim et al., 2018; Anderson et al., 2017; Fukui et al., 2016) to fuse image representations with questions for VQA models, most models only use simple RNNs such as LSTM(Hochreiter and Schmidhuber, 1997) or GRU(Cho et al., 2014) when encoding the input question. The main reason is that RNNs encode the words in the question sequentially to represent the sentence structure. However, for the VQA models, (Agrawal et al., 2016) showed that VQA models strongly rely on language bias and often jump to conclusions after seeing the first several words in the question. Furthermore, (Mudrakarta et al., 2018) showed that existing VQA models depends on only a little parts of the question words to answer the question. Also, (Kafle and Kanan, 2017) and (Kazemzadeh et al., 2018) investigated the inconsistency of existing VQA models. Those results explains that VQA models may not utilize the sentence structure in questions and just rely on specific words in the question.

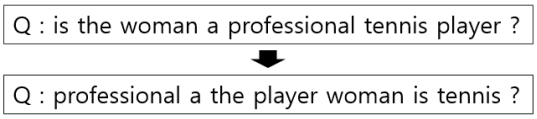

Figure 1: Shuffling the Word Order of Question

For other vision-language co-understanding tasks named Referring Expression Recognition(RER)(Kazemzadeh et al., 2014), (Cirik et al., 2018) showed that randomly permuting the word order of expression has little impact on the recognition model.

In this paper, we study the weakness of question understanding in existing VQA model specially for insensibility of question word order like (Cirik et al., 2018) did in RER. The observation of our study is the following.

1. We observe that shuffling the word order of input question randomly to destroy the sentence structure has little impact on existing VQA model.

2. We demonstrate that encoding the question with self-attention based architecture which does not consider word order, shows almost same performance to current RNNs architecture's performance in existing VQA model.

2 Experiments

Baseline Model

We use bottom-up top down attention model(Anderson et al., 2017), the winning model for 2017 VQA challenge, as a baseline model for our experiments. The model encodes the questions with RNNs and then adapt soft

Accuracy	Train-Org	Train-Rand
Eval-Org	63.25	60.91
Eval-Rand	49.26	60.98

Table 1: Random Permutation Results. Org means that the input is in the original order and Rand is in the randomly shuffled order while it is trained or evaluated.

attention mechanisms with the image features extracted from faster RCNN(Ren et al., 2017).

Implementation Details

We use 1-layer bi-directional GRU with 1024 hidden units to encode the question and use 2-layer feed-forward neural networks to final layer. We use pre-trained GloVe(Pennington et al., 2014) word embedding and freeze during the training. We use Adam Optimizer(Kingma and Ba, 2014) for learning late to 0.001 and train 30 epochs.

Dataset

All experiments in our paper are conducted on VQA v2 dataset(Goyal et al., 2017). We trained models on train split with 82,783 images and 443,757 questions and evaluate on validation set with 40,503 images and 214,354 questions.

2.1 Shuffling the Word Order of Question

To check the whether the model utilizes the question word order with RNNs in visual question answering, we randomly shuffled the question words like Figure 1 in the question. We report the result in Table 1 for three cases, shuffling for both train set and evaluation set, only for train set and only for evaluation set. As shown in Table 1, the model trained with word shuffling for both of the splits and only for train splits show slight performance drop. One of the reasons is that the average question length of the VQA dataset is short to need reasoning. The other reason is that the model mostly uses the information from specific words as shown in (Agrawal et al., 2016; Mudrakarta et al., 2018). Unlike the case that is only shuffled for evaluation set, shuffling the word order during the training strengthens this phenomenon. Hence, unless the specific words are dropped, changing the order of the words may not hurt the question representation much.

2.2 Self-Attention Mechanism

We conduct experiments about the models other than RNNs that does not consider the word order

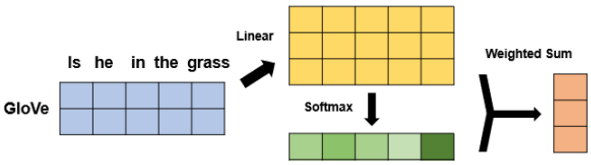

Figure 2: Self-Attention Mechanism

Question Encoding	Accuracy
GRU	63.25
Average Pooling	61.01
Self-Attention	63.06

Table 2: Comparison with Order-Independent Methods

of input question. At first, we simply substitute RNNs in base model with average pooling of word embedding to represent the question. Since this model does not receive the position of the word in the question as an input, this model is independent of question word order. As shown in Table 1, the performance drops slightly from original GRU based model and shows similar performance to word shuffled RNNs. Also, we substitute RNNs in baseline model with self-attention mechanism. This model, as shown in Fig 2, encodes a input question with a linear transformation, self-attention layer and weighted summation of input word embedding. This model is also independent of question word order. As shown in Table 2, this model shows almost same performance as RNNs. However, since it learns the importance of specific words in the sentence, the model results in better performance than average pooling. This results seems that word order is unnecessary feature for existing VQA models.

3 Conclusion and Future Work

In this paper, we observe that shuffling the word order of input questions in current VQA model has little impact on the performance even if the model use RNNs to encode the question sequentially. Also, we observe that encoding input questions with simple self-attention based network that is independent of word order, show almost same performance with RNNs. Hence, we observe that VQA models does not utilize the question word order to answer the question. Based on those observations, our future work will be about adopting new encoding methods to VQA models such as Graph Convolutional Networks(Defferrard et al., 2016) to better understand the input question.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1955–1960.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998.

Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: multimodal tucker fusion for visual question answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2631–2639.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 781–787.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3837–3845.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 457–468.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2018. Make up your mind: Towards consistent answer predictions in vqa models. In *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, pages 124–132.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 1571–1581.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1896–1906.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.